**CREDITBANK of Australia**

**Virtual Data Analysis Case Study**

**Background Introduction:**

CREDITBANK of Australia is one of the major banks in Australia with offices in all the state capitals and offers a wide range of personal as well as business banking products & services to their retail and business customers. It provides a variety of loan services such as personal loans, home loans, business loans, auto loans, study loans, and so on. The Bank operates on a slim profit margin as it offers attractive interest rates on fixed deposits as well as loans. As a result, CREDITBANK of Australia leads the Australian market as one of the largest loan providers that the customers rely on.

The default loans, also known as Charged Off Loans, are usually declared when payment is 270 days late. Banks can lose a huge sum of money when borrowers default, which in turn, forces banks to reassess their lending procedure to reduce their credit risk. In recent years, loan default rates have increased steadily in the Australian market. Compared to other banks, thanks to its strategy of operating with a slim profit margin, CREDITBANK of Australia is more vulnerable to the impact of increasing loan default rates. The management believes that building analytics capability to understand and predict future loan defaulting applications can help the managers to make better decisions on future loan applications. The management has already started gathering data from both internal and external sources to build models that help to make informed decisions on loan applications.

**Main Task:**

Firstly, as a data analyst of CREDITBANK of Australia to develop two or more loan default prediction models and select the model that is more appropriate for this case and explain the reason. Then, present it to the executives who would make the decisions on whether to implement it or not.

1. Executive Summary:

A short summary of the key information in the document and is written for a businessperson. It should contain clear, well-structured, evidence-based summary of your findings from your models and analysis and should provide insights for CREDITBANK of Australia managers on what to consider in order to effectively address the problem of customers defaulting loan application.

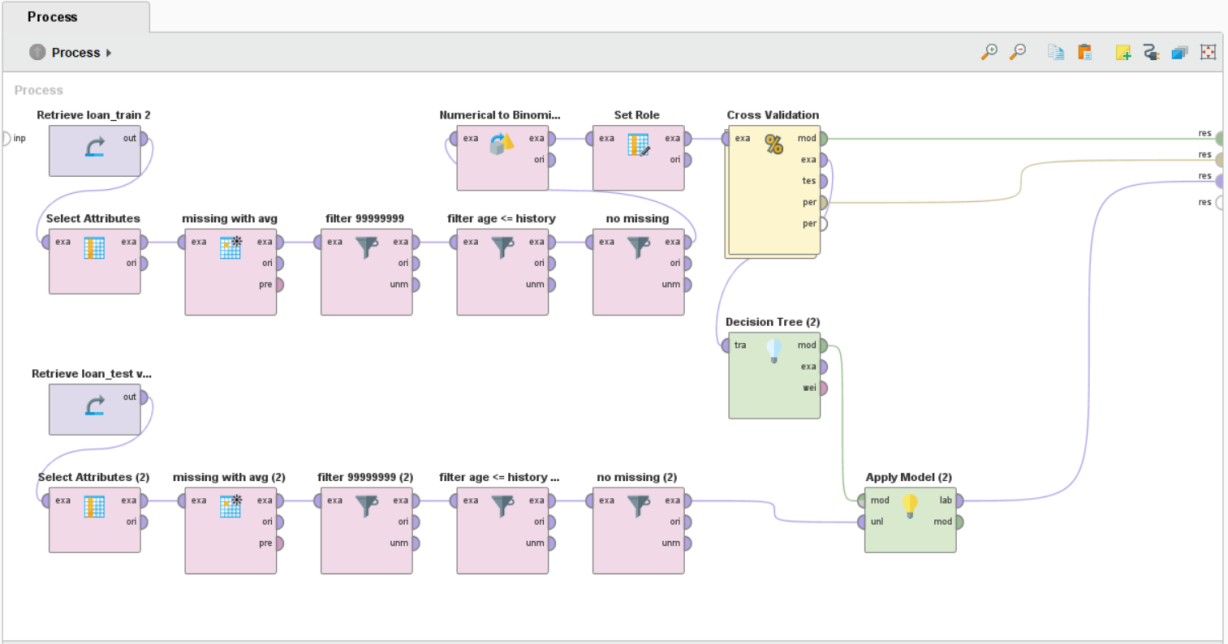
This report is based on the personal information collected from participants and analyses and discusses what factors are determinants of defaulted loans. The author first understand and pre-process the given data sets of the various attributes and specific values. Based on ethical considerations, the author pre-processed the data to remove some of the attributes that could lead to discrimination when reading the reports and obtained insights.

In order to verify the accuracy of the analysis and predictions, a decision tree model and a logistic regression model were then used and the decision was made to use the decision tree model that produced the more convincing results based on their respective performance and a review of the relevant literature to analyse and discuss the results.

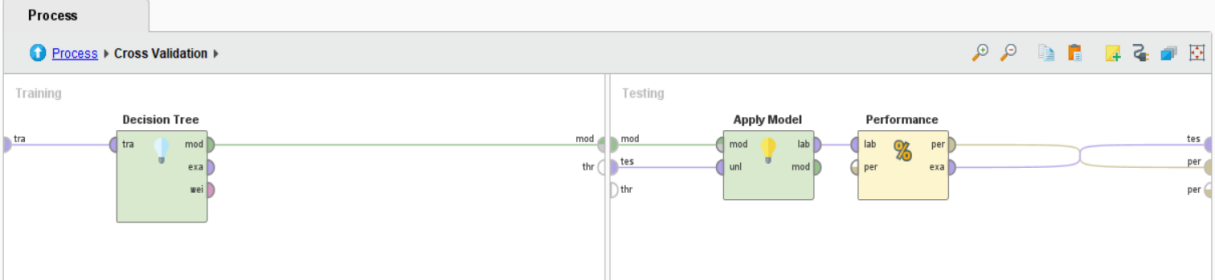
By analysing and researching the personal information of the borrower in advance, bank could effectively resolve the default of the customer. Based on the analysis of the decision tree visualisation, it can be seen that the user's disposable cash flow (income level, bank savings) and the ratio of capital income to capital loss are decisive factors in predicting whether a lender will be trustworthy. Customers with healthier assets tend to have a better capability to repay debts and thus not default on their loans. In addition, it is also relatively safe for bank to lend to customers with a lower cumulative current loan size and a higher level of education. In addition, customers of a similar age who have been applying for and using credit cards for longer period tend to have better credit rates and are more likely to pay loans on time.

# Predictive analysis of loan default via RapidMiner:

* + 1. 1) **Screenshot** of Analytical Process (first model: decision tree)

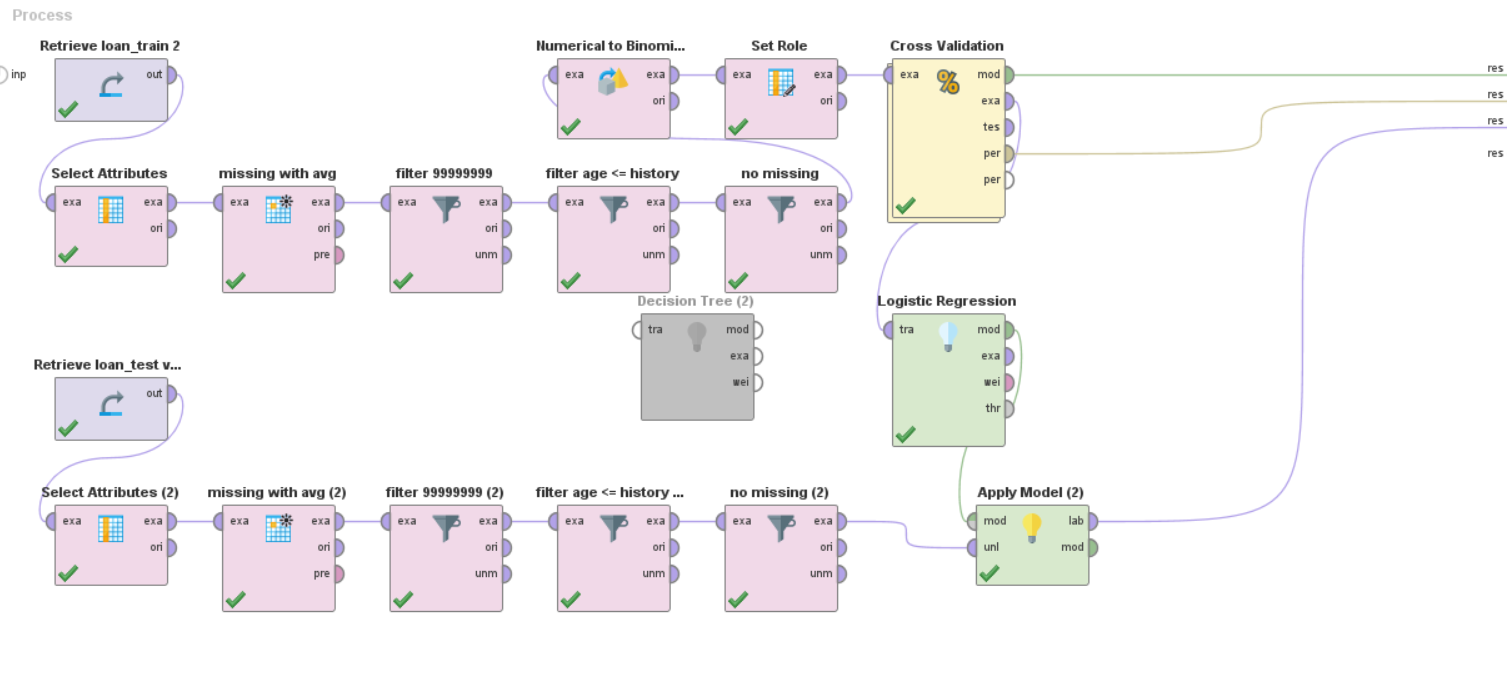


\* Figure 1: Overall process

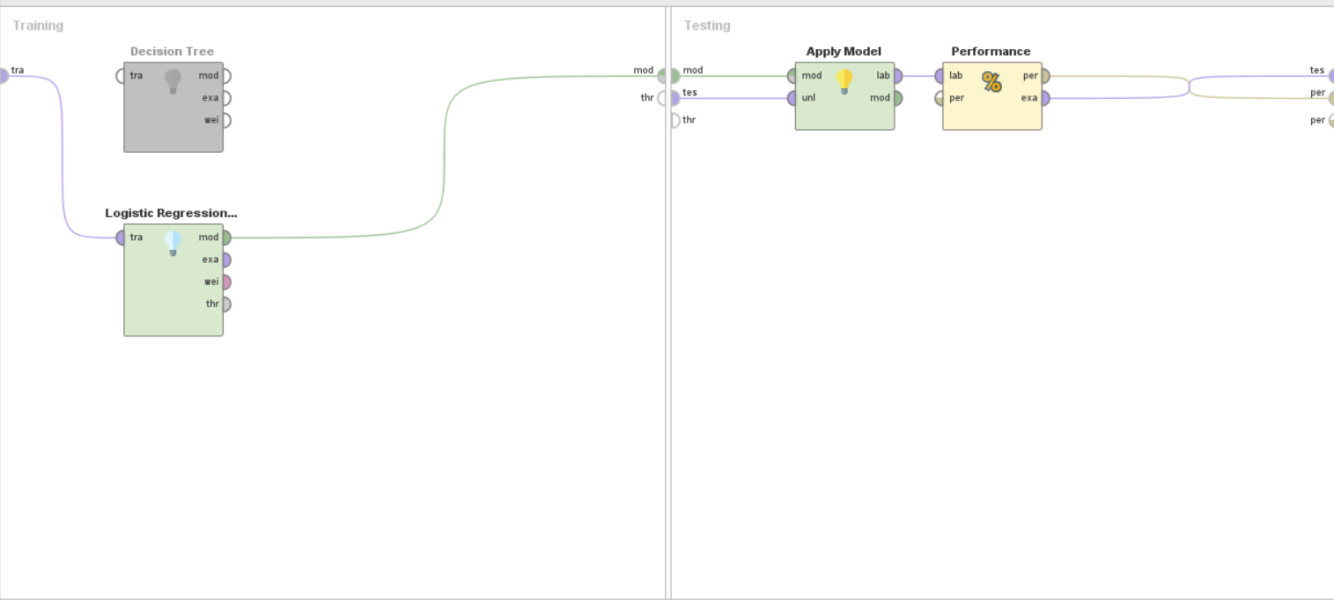


\* Figure 2: Process within Cross Validation

1. **Explanation:** Firstly, 15 variables are selected with enough consideration (detailed information please refer to 3.1.2). The missing data for age were replaced by its average value. As for the unusual value for attribute loan\_amount is 99999999, those data were filtered out. The credit history of a person cannot exceed its age, thus there is another operator to filter out the rows which have a greater value for credit history than for age. As for remaining missing data, considering the small amount of these data in the overall dataset, the choice was made to remove them directly. Then, process the data to transfer the data type of default from numerical value to binominal value and set it as the label. Also, set attribute id as the index. Then, model training and integration with the test data set can be implemented. Within Cross Validation, the decision tree is selected to the training sub-process (please refer to Figure 2).
   * 1. 1) **Screenshot** of Analytical Process (second model: logistic regression)



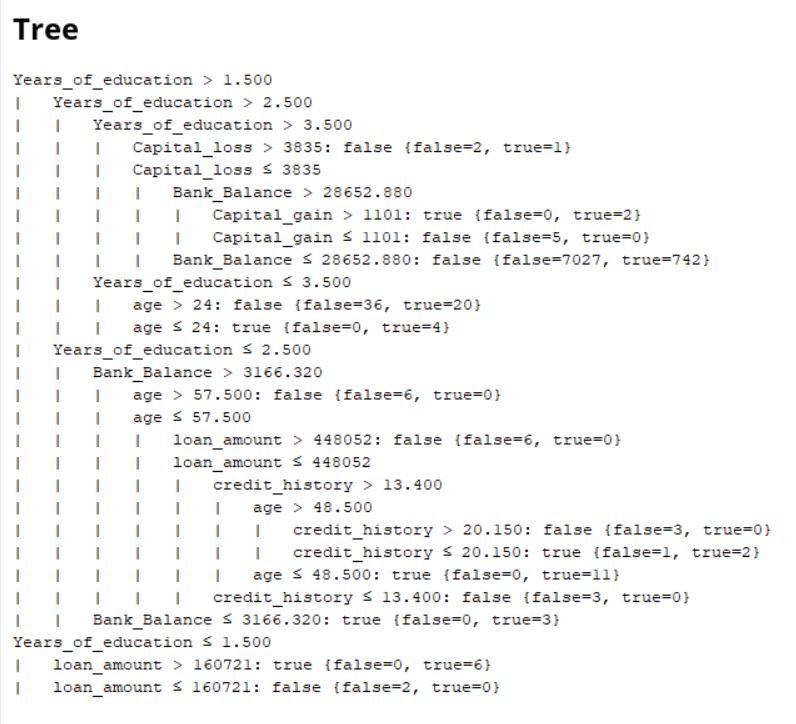
\* Figure 3: Overall process



`

\* Figure 4: Process within Cross Validation

1. Explanation the most decisive predictor(s) of loan default regarding each model
   * 1. Screenshot of Results of the first model (decision tree)

****

\* Figure 5: Results of first model (decision tree)

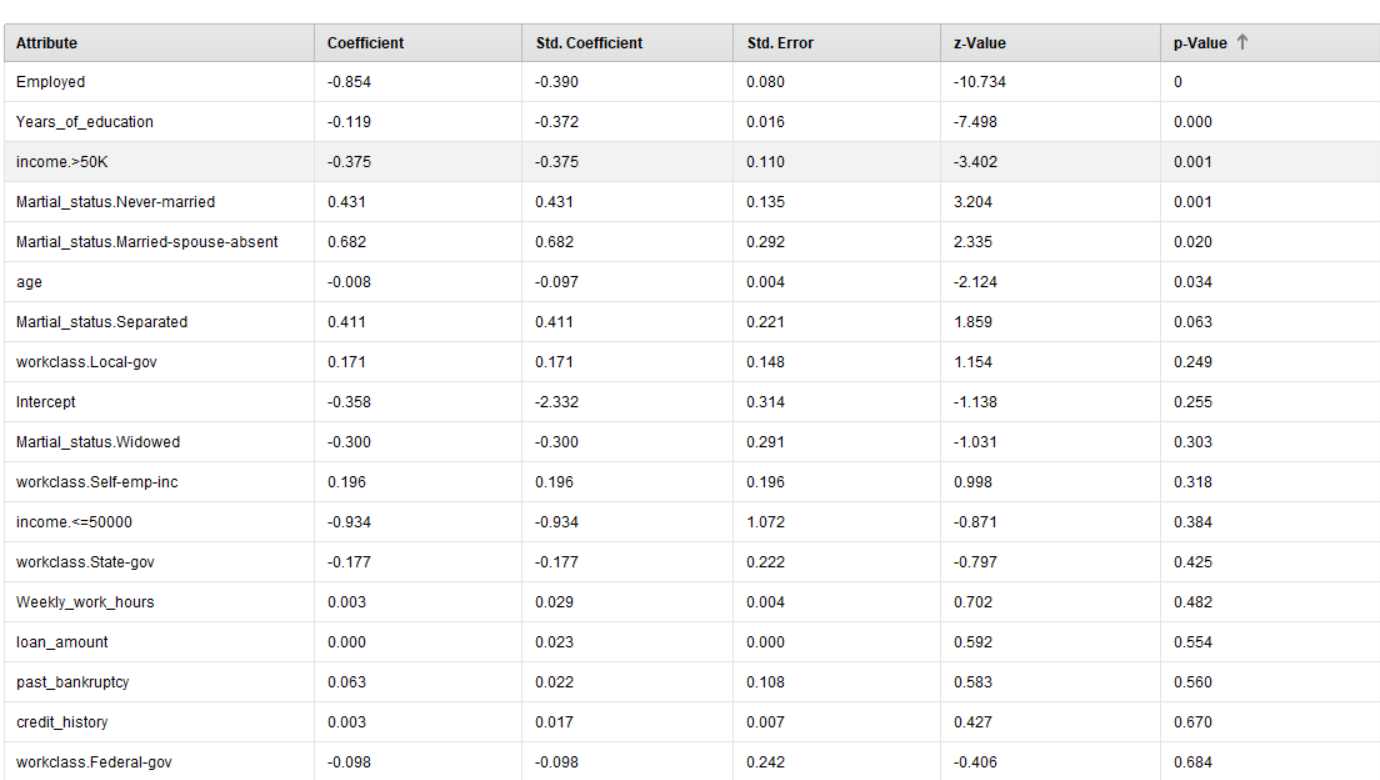
* + 1. Interpretation of the results of the first model (decision tree)

*(Explain the choice and impact of the most decisive predictors)*

When considering which variables should be selected, attribute Education can be described as the variable Year\_of\_education. By looking into the data information, the higher people's academic qualifications means that they will take longer to study. It is clearly redundant when both parameters in the data set are able to indicate the educational level of the respondent. Based on the common sense, Variable loan \_purpose was not selected due to the limited relevance it have to the result. When considering the ethical implications, the following attributes Gender, Country/area\_of\_origin, Occupation and Race were not selected, as those attributes is likely to generate discrimination against people in varying degrees.

According to the description of the decision tree above, there are seven decisive predictors among the selected variables. Attribute Years\_of\_education is the most decisive predictors, which is used as the prerequisites (largest impact) to determine and analyse different cases. Attributes Capital\_loss, Bank\_balance, Capital\_gain, age, credit\_history, loan\_amount are also the decisive predictors.

* + 1. Screenshot of results of the second model (logistic regression)





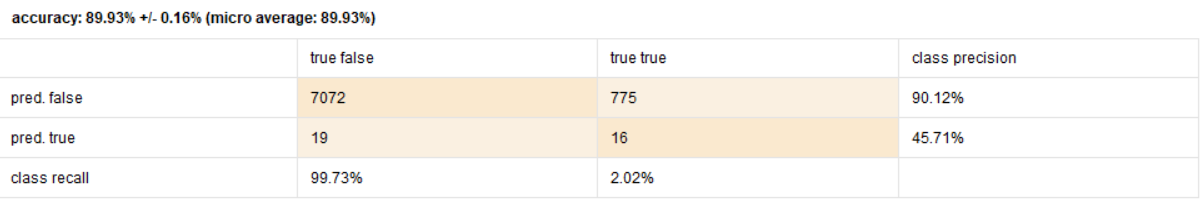
\* Figure 6: Results of second model

* + 1. Interpretation of the results of the second model (logistic regression)

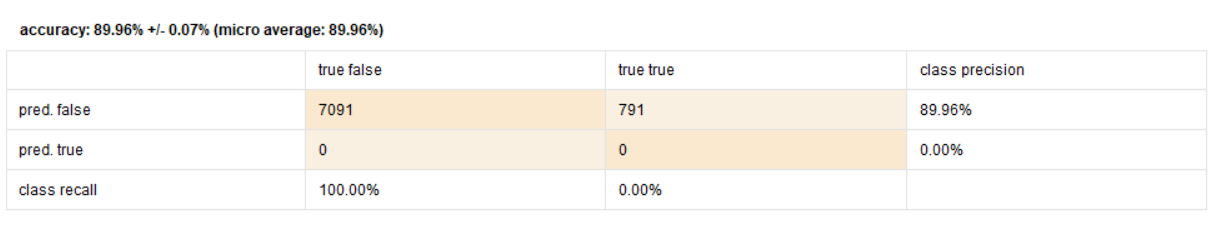
*(Explain the choice and impact of the most decisive predictors)*

The second model has the same principle of selecting variables as in the first model. In comparison, In the case of logistic regression, the relationship between the variables and the hypothesis is judged on the basis of p-value. If the p value is smaller, this means that there is a higher chance of a relationship between the two variables [1]. In descending order of p-value, employment status, education (Year\_of\_education), income (income >= 50K, income < 50K), type of work (Local-gov, Self-emp-inc, State-gov), marital status (never-married, married- spouse-absent , separated, widowed), age, weekly\_work\_hours are all decisive factors in predicting the creditworthiness of a borrower and and all determine the accuracy of the prediction to varying degrees.

1. Compare the model performance of both models and justify which method should the management use:
   1. Screenshots of Results (both models)



\* Figure 7: Performance of first model (decision tree)



\* Figure 8: Performance of second model (logistic regression)

* 1. Interpretation and justification of model performances

*(Describe the model performances and describe which model is better for this analytical task)*

The performance of the two models is similar in terms of numbers alone, each with an accuracy rate of around 89%. Decision tree model is still a better option to do this analytical task for the following reasons.

When reviewing and analysing the loan\_train dataset used to train the model, it is easy to see that the missing values and outliers were filtered out manually during the data pre-processing process. These filtered values may affect the accuracy of the originality of the data set and if these outliers are failed to detect during the pre-processing of the data, the model using logistic regression is more likely to suffer even more. In addition, the decision tree is a graphical visualisation that is more accessible to those who read the predictions.It is easier for bank to judge different borrowers based on the criteria displayed in the decision tree, for example the two outermost branches of the decision tree are separated by whether the customer has more or less than 1.5 years of education. In the second model, although it is clear that the length of education is a determinant of the borrower's compliance, it is not clear in what range of education the borrower is more likely to default on the loan. This is because logistic regression is more applicable to continuous data and is not as useful in situations where the predicted outcome needs to be discussed in a disaggregated manner [2].

1. **Generate Business Insights and Strategize: Generate Business Insights and Strategize:**In this reflection, the business knowledge should also be applied:
   1. Offer compelling explanations for the observed findings (e.g., regarding the financial health and employment information) and suggestions for additional analyses (or data) that could be used to investigate them.

*(Develop explanations for the reported findings and describe means on how to potentially analyze and improve the data or model or analysis)*

Based on Figure 5, the results of the decision tree indicate that the education level of the participants is the most critical of all the selected attributes in predicting the creditworthiness of the lender. When considering the relationship between academic qualifications and employment, it is commonly believed that the higher levels of education can lead to higher paying jobs based on fewer security risks [3, 4]. Also, the level of income will directly affect the amount of savings [5]. In addition to disposable cash flow, the ratio of capital gain to capital loss is also a crucial factor to determine the level of lender’s revenue [6].The amount of loan is an important factor in judging the health of a property [7] and the financial health could directly determine whether the loan will be defaulted.

When considering how to improve the model to make the data analysis more accurate, a combination of decision trees and logistic regression would be a better alternative [8]. This makes better sense for use with data that has not been pre-processed in detail.

* 1. Propose specific actions on how to address the identified factors that contribute to loan default.

*(Describe means on how to potentially overcome the loan default-drivers that have been Identified)*

According to the results of the preferred models developed with the decision tree above, it is relatively safer to lend to people with a healthy financial status. A high level of income and positive capital gain could be considered as a healthy asset state [7], which tends to indicate a better ability to repay the loan. In addition to this, customers with a great credit rating and a small amount of current loan arrears are more likely to be able to pay their loans on time. For lenders of a similar age, the shorter the period of education might result in the greater the probability of default on a loan.

* 1. Describe the potential ethical and the legal issues that the bank needs to be concerned of.

Although some of the attributes that would cause banks to discriminate (racial discrimination, regional discrimination, gender discrimination, specific occupational discrimination) when reviewing information about lenders have been eliminated in the modelling process. However, there are still some potential ethical issues. This includes age discrimination against relatively vulnerable groups of older people in society.

In the results of the case study using decision trees, it is clear that some of the researcher's own subjective judgements may not be decisive in predicting whether the lender will be late in repaying the loan. Those factors that are not decisive at all contains a great deal of private information about the participants and some of the sensitive content that they do not want to disclose for various reasons (e.g the number of times the lender has been insolvent and the number of hours worked per week.). Although the inclusion of comprehensive information in the dataset helps the researcher to conduct the study, such an overly intrusive way of analysing the details of the participants' private lives can certainly make them feel anxious about it.

When a bank conducts a research case based on a participant's personal information, the disclosure of the user's private information for technical reasons or due to ethical insensitivity or neglect is a violation of the Australian data privacy Laws [9]. When collecting and using participants' personal information, banks must reach agreements with them that enable them to solicit private information and fully respect and accommodate participants' individual requests during research the process (e.g. in the form of anonymity in the data set) [10].

In conclusion, the research is conducted in a manner that respects the individual rights of the participants and does not compromise their dignity and interests [11].

**References**

[1] "How do I interpret p value in logistic regression?" <https://www.quora.com/How-do-I-interpret-p-value-in-logistic-regression> (accessed.

[2] "Logistic Regression vs. Decision Tree." <https://dzone.com/articles/logistic-regression-vs-decision-tree#:~:text=%20Logistic%20Regression%20assumes%20that%20the%20data%20is,not%20sure%2C%20then%20go%20with%20a%20Decision%20Tree>. (accessed.

[3] "3 ways education impacts health." <https://healthyopportunitiesin.com/3-ways-education-impacts-health/> (accessed.

[4] "Education as a Social Determinant of Health." <https://publichealth.tulane.edu/blog/social-determinant-of-health-education-is-crucial/#:~:text=Not%20only%20does%20education%20give%20individuals%20a%20chance,about%20how%20to%20take%20care%20of%20their%20health>. (accessed.

[5] T. Pettinger, "Factors that influence saving levels," 2021. [Online]. Available: <https://www.economicshelp.org/blog/146244/economics/factors-that-influence-saving-levels/#:~:text=Quick%20Summary%20of%20Factors%20That%20Influence%20Saving%20Levels>.

[6] K. Kiyosaki, "Cash Flow vs Capital Gains: Understanding the 2 Types of Investment Income," 2022. [Online]. Available: <https://www.richdad.com/cash-flow-vs-capital-gains>.

[7] J. Kagan. "Financial Health." <https://www.investopedia.com/terms/f/financial-health.asp> (accessed.

[8] "Combining logistic regression and decision tree." <https://towardsdatascience.com/combining-logistic-regression-and-decision-tree-1adec36a4b3f> (accessed.

[9] "Australian Data Privacy Laws to Know in 2022." <https://www.groundlabs.com/blog/australian-data-privacy-laws-to-know-in-2022/> (accessed.

[10] "Overview of Privacy Law in Australia." <https://hallandwilcox.com.au/thinking/overview-of-privacy-law-in-australia/> (accessed.

[11] J. Mandal, S. Acharya, and S. C. Parija, "Ethics in human research," *Tropical parasitology,* vol. 1, no. 1, p. 2, 2011.